Dual Diffusion for Unified Image Generation and Understanding

Supplementary Material

1. Training Details

Hyperparam.	Dual pretrain	Continued pretrain		Instruct. tuning
		Mask emb.	High res.	
Gradient steps	60k	200k	80k	50k
Batch size	512	512	768	512
LR	5e-5	3e-5	3e-5	3e-5
Scheduler	Constant LR with warmup			
Warmup iters	5000	1000	1000	1000
Weight decay	1e-2			
Text loss weight	0.2			1.0

Table 1. Training hyperparameters for D-DiT. Text loss weight denotes the λ in Equation (8).

We provide the detailed hyperparameter setting for different training stages in the Table 1. During all the training stages, we use AdamW optimizer with default hyperparameters ($\beta_1 = 0.9, \beta_2 = 0.999$). Mixed precision training (bf16) and fully-sharded data parallel (with gradient and optimizer state sharded) are used for model training.

2. Further Results

Model	Backbone	Params. (B)	FID ↓
SD-XL [8]	Diff.	0.9	9.55
PixArt-α [2]	Diff.	0.6	6.14
Playground v2.5	Diff.	-	4.48
Show-O [10]	Discrete Diff.	1.3	15.18
LWM [7]	AR	7	17.77
VILA-U [9]	AR	7	7.69
SD3 [3]	Diff.	2	16.45
D-DiT	Diff.	2	15.16

Table 2. Comparison with other models on MJHQ-30K evaluation benchmark at 512×512 resolution.

Model	COCO-30k		T2I CompBench		
	FID ↓	CLIP ↑	Color ↑	Shape ↑	Texture ↑
SD3	10.2	30.9	0.7993	0.5816	0.7389
D-DiT	9.4	31.2	0.8001	0.5703	0.6856

Table 3. Further image generation comparisons against original SD3 on MS-COCO dataset [6] and T2I CompBench [4].

Image generation We evaluate the aesthetic quality of generated images from our proposed D-DiT against those of the original SD3 model and a selection of existing text-to-image (T2I) and multi-modal works. We measure Frechet Inception Distance (FID) with respect to a collection highly aesthetic generated images, known as the MJHQ-30K benchmark proposed by [5]. As shown in Table 2, we observe an improvement in FID after joint diffusion training, and favorable comparison against multi-modal models of similar size. We also provide further comparisons on MS-COCO 30k and T2I CompBench in Table 3. The FID and CLIP score slightly improve compared to the original SD3 model. On T2I CompBench, we find that after dual diffusion fine tuning the model performs worse in texture. We hypothesize that the major reason is the texture quality of our training dataset is worse than the dataset used for training SD3.

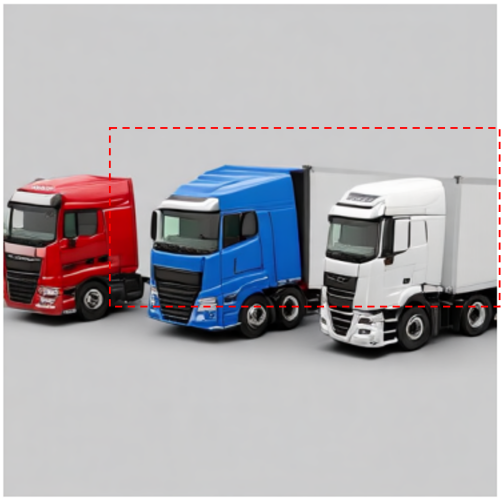
Text generation process We provide an illustrative example of masked diffusion in Figure 2 for the visual question answering task, where the token generation process is visualized over diffusion time. Over the course of sampling, the answer tokens are gradually denoised from the masked state via absorbing state reverse diffusion. The question tokens are always left unmasked throughout the entire process.

Model	# trainable	Text encoder	Geneval	COCO FID	VQAv2(val)	
					0-shot	32-shot
End-to-End	1.1B	-	0.39	18.1	54.3	58.7
From SD3	2B	T5-XXL	0.65	9.4	55.0	60.3

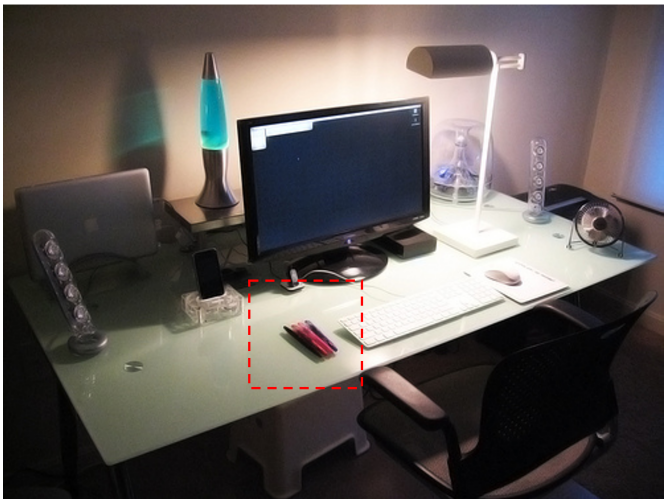
Table 4. Comparison of different D-DiT variants. *End-to-End* variant is trained from scratch and uses GPT2's text tokenizer. *From SD3* variant is initialized from SD3 pretrained checkpoint and uses T5 encoder. The end-to-end model is first trained on OpenWebText for 350B tokens, then trained on DataComp-recap1B for an epoch (400k steps) and a filtered subset for 100k steps.

Training from scratch and removing T5 encoder To study the influence of text-to-image pretraining, we conduct a study by comparing a D-DiT model that is trained from scratch. We found that initializing from pretrained text-to-image model and use a pretrained text encoder can greatly aid model learning of text-to-image tasks. Meanwhile, image captioning on VQA also mildly improves (Table 4).

Image generation's influence on SFT To analyze the influence of dual diffusion loss on image understanding, we conduct supervised finetune on LLaVA 1.5 dataset with



(a) T2I Prompt: *Three trucks parking in parallel: one red, one blue, and one white. Red truck has load and the rest don't have.*



(b) I2T Prompt:*Q: How many pens are there on the desk and what are their colors?* A: There are three pens on the desk, and they are red and blue.

Figure 1. Examples of failed text-to-image and image-to-text generation.

varying amount of image generation data, including a training that only has understanding loss (no generation data). We observe that the image generation loss and corresponding data amount does not have significant influence on model's understanding performance (Table 5).

Und.	Gen.	VQAv2 (val)			POPE		
		10k	30k	50k	10k	30k	50k
0.665M	0	52.8	55.9	58.3	79.6	80.9	81.8
0.665M	7M	53.4	55.8	58.1	79.8	81.2	82.4
0.665M	20M	53.6	55.8	58.3	81.0	81.1	82.5

Table 5. Understanding performance (accuracy) under different data settings and training steps during supervised finetune. Batch size is set to 128 for this experiment.

Comaprison against previous multi-modal diffusion model We also include a qualitative comparison in captioning performance compared to UniDiffuser [1], another diffusion-based multi-modal model, in Figure 3, where we demonstrate an improvement in the ability to capture fine-grained details of the image in a longer caption format. Finally, we provide further uncurated text-to-image (T2I) generation results in Figures 4, 5, 6, and 7. Overall, these results further demonstrate the multi-faceted performance of our proposed dual-branch diffusion-based multi-modal model.

Limitations As shown in Figure 1b: in T2I, we find that D-DiT can struggle to generate scenes with relatively complex instructions. In I2T, D-DiT can fail to identify the full details of smaller objects. We also observe model's performance performance deteriorates with longer prompts, primarily due to the bias towards short prompts in the LLaVA finetuning dataset.

In summary, while discrete diffusion offers the advantage of being agnostic to sequential order and is compatible with bi-directional Transformers, its current implementation requires the sequence length to be preset before sampling. A promising future direction would be to extend the sampling scheme to allow for more flexibility, enabling dynamic sequence lengths during the sampling process. In addition, while we show that our proposed dual diffusion model can perform instruction tuning, its instruction-following capabilities still marginally lag behind those of state-of-the-art autoregressive models.

t=1

Q: Provide a brief description of the given image. A: [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
[MASK] [MASK] [MASK] [MASK] [MASK] [MASK [MASK] [MASK] [MASK]

Q: Provide a brief description of the given image. A: [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] beautiful [MASK] [MASK] taken
[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] to [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
[MASK] [MASK] [MASK] of [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] town [MASK] [MASK] [MASK] [MASK] [MASK]
[MASK] [MASK]. [MASK] [MASK]' [MASK] [MASK]ray [MASK] illuminate [MASK] [MASK] [MASK] [MASK] [MASK]a warm [MASK] [MASK] [MASK]
[MASK] [MASK]zure [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]</s> [MASK]</s> [MASK] [MASK]
[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]</s> [MASK] [MASK] [MASK] [MASK] [MASK]
[MASK]</s> [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]</s></s> [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]</s></s> [MASK]

Q: Provide a brief description of the given image. A: [MASK] image [MASK] a serene and beautiful [MASK] scene taken [MASK] [MASK] [MASK]
[MASK] [MASK], which appears to be a [MASK] [MASK] [MASK] From [MASK] wooden balcony [MASK] [MASK] of a clear [MASK] [MASK] where
the snowy town [MASK] the base [MASK] [MASK] horizon. [MASK] [MASK]' [MASK] [MASK]ray [MASK] illuminate the landscape [MASK] casting
[MASK]a warm glow [MASK] the azure backdrop [MASK]</s></s></s> [MASK] [MASK]</s></s> [MASK]</s></s></s></s></s></s></s></s>
[MASK]</s></s> [MASK] [MASK]</s> [MASK] [MASK]</s> [MASK]</s></s></s> [MASK]</s> [MASK]</s></s> [MASK]</s></s></s></s> [MASK]
[MASK]</s></s></s> [MASK] [MASK]</s> [MASK]</s></s></s></s>

Q: Provide a brief description of the given image. A: [MASK] image [MASK] a serene and beautiful [MASK] scene taken [MASK] [MASK] [MASK]
[MASK] [MASK], which appears to be a [MASK] [MASK] [MASK] From [MASK] wooden balcony [MASK] [MASK] of a clear [MASK] [MASK] where
the snowy town [MASK] the base [MASK] [MASK] horizon. [MASK] [MASK]' [MASK] [MASK]ray [MASK] illuminate the landscape [MASK] casting
[MASK]a warm glow [MASK] the azure backdrop [MASK]</s></s></s> [MASK] [MASK]</s></s> [MASK]</s></s></s></s></s></s></s></s>
[MASK]</s></s> [MASK] [MASK]</s> [MASK] [MASK]</s> [MASK]</s></s></s> [MASK]</s> [MASK]</s></s> [MASK]</s></s></s></s> [MASK]
[MASK]</s></s></s> [MASK] [MASK]</s> [MASK]</s></s></s></s>

Q: Provide a brief description of the given image. A: The image presents a serene and beautiful winter scene taken from a vantage point, which appears to be a mountain range. From the wooden balcony in front of a clear blue sky where the snowy town at the base meets the horizon. The sun's rays illuminate the landscape, casting a warm glow against the azure backdrop.
</s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s><
/s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s></s>

t=0

Figure 2. Illustrative example of visual question answering with mask diffusion.

D-DiT: The image features a phone held up in an interesting angle, standing on a surface.

UniDiffuser: a white iPhone sitting on top of a stand

D-DiT: In the image, the saucer and cup is laid horizontally on one of the mats.

UniDiffuser: A set of three blue and white striped napkins

D-DiT: The image shows a woman walking down a runway in her model outfit. The outfit includes a coat, a book, a skirt, and a purse or handbag. She is also wearing tall boots.

UniDiffuser: A model walks down the runway in a beige coat and boots

D-DiT: In the image, there are three baseball players, all of which are all dressed in white uniforms. The first man appears to be cheering to hit the ball. The other two players, possibly his teammates or fielders, are in different positions on the field.

UniDiffuser: Jonny Bairstow of Australia celebrates after taking the wicket

D-DiT: The image captures a captivating view of a outdoor concert with a glow of night. The concert is taking place at dusk and features a large stage with colored purple lights, creating a stunning visual and vibrant setting. A crowd can be seen sitting around the area, enjoying the musical performance on the stage. The balkan-ish skies of the evening sunset adds warmth to the scene, further enhancing the concert atmosphere.

UniDiffuser: A large crowd of people on stage at a concert

Figure 3. Comparison of captions generated by D-DiT and UniDiffuser[1]. The prompt to D-DiT is "Provide a brief description of the given image."

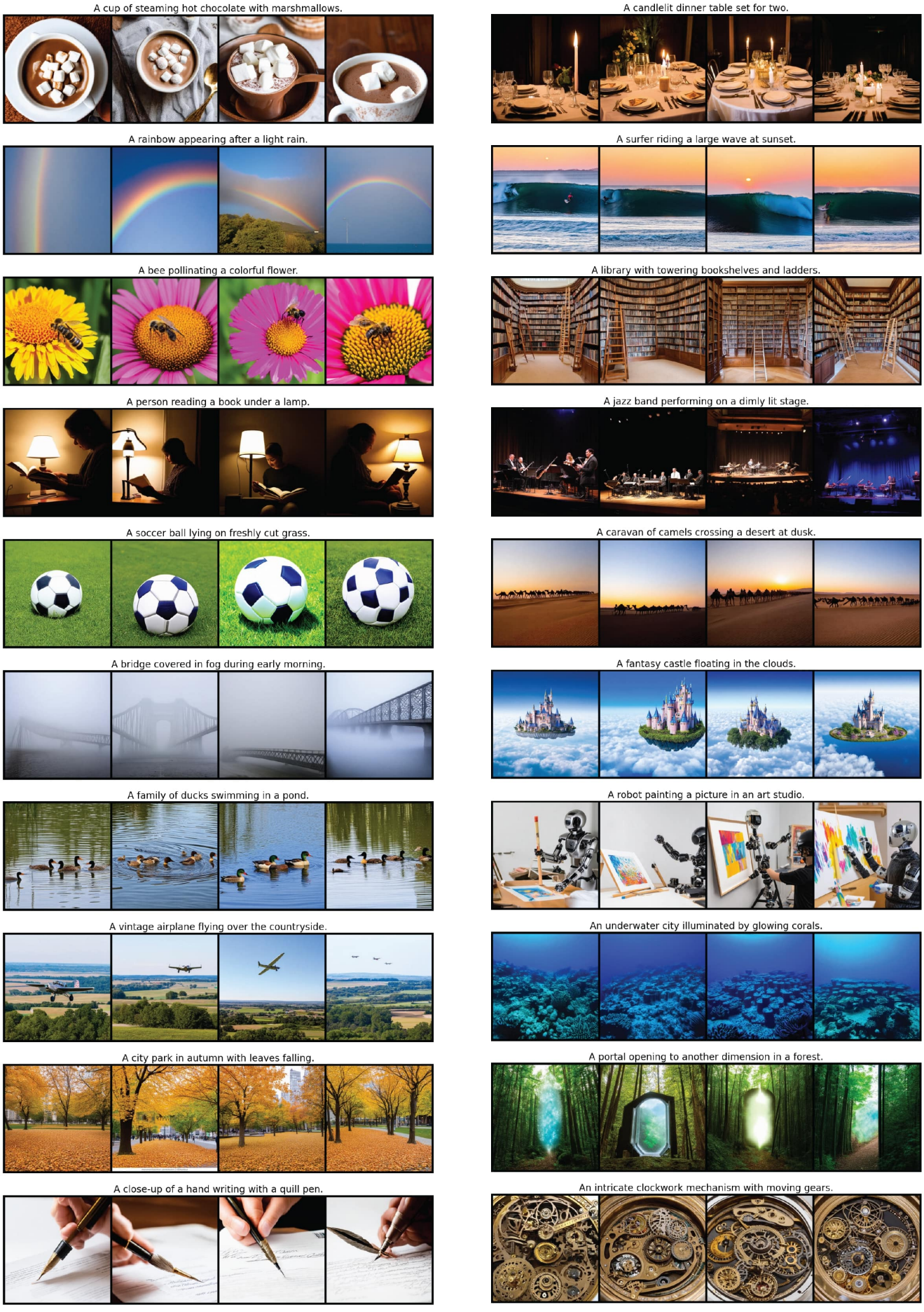


Figure 4. Additional text-to-image samples generated from the model.

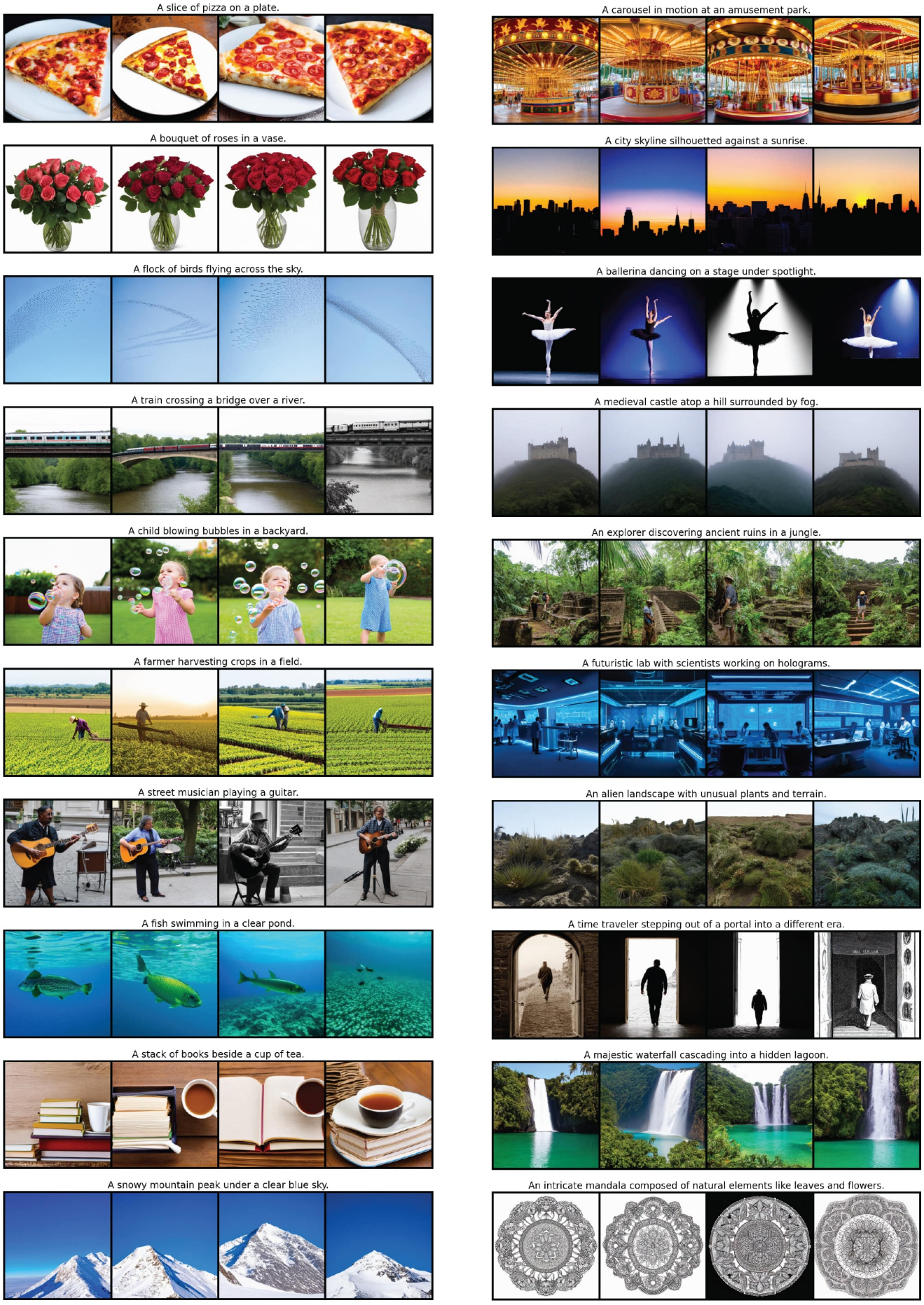


Figure 5. Additional text-to-image samples generated from the model.

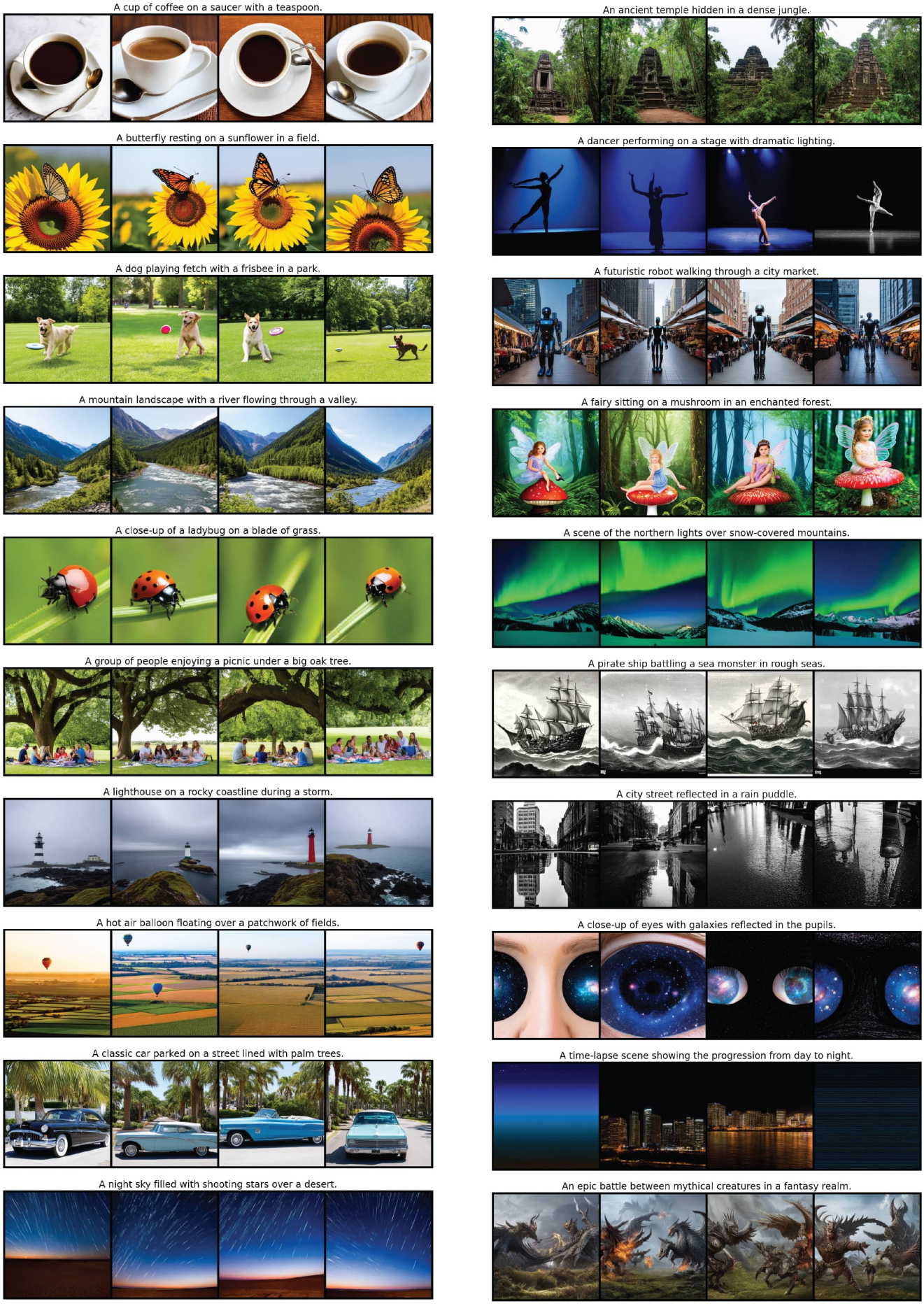


Figure 6. Additional text-to-image samples generated from the model.

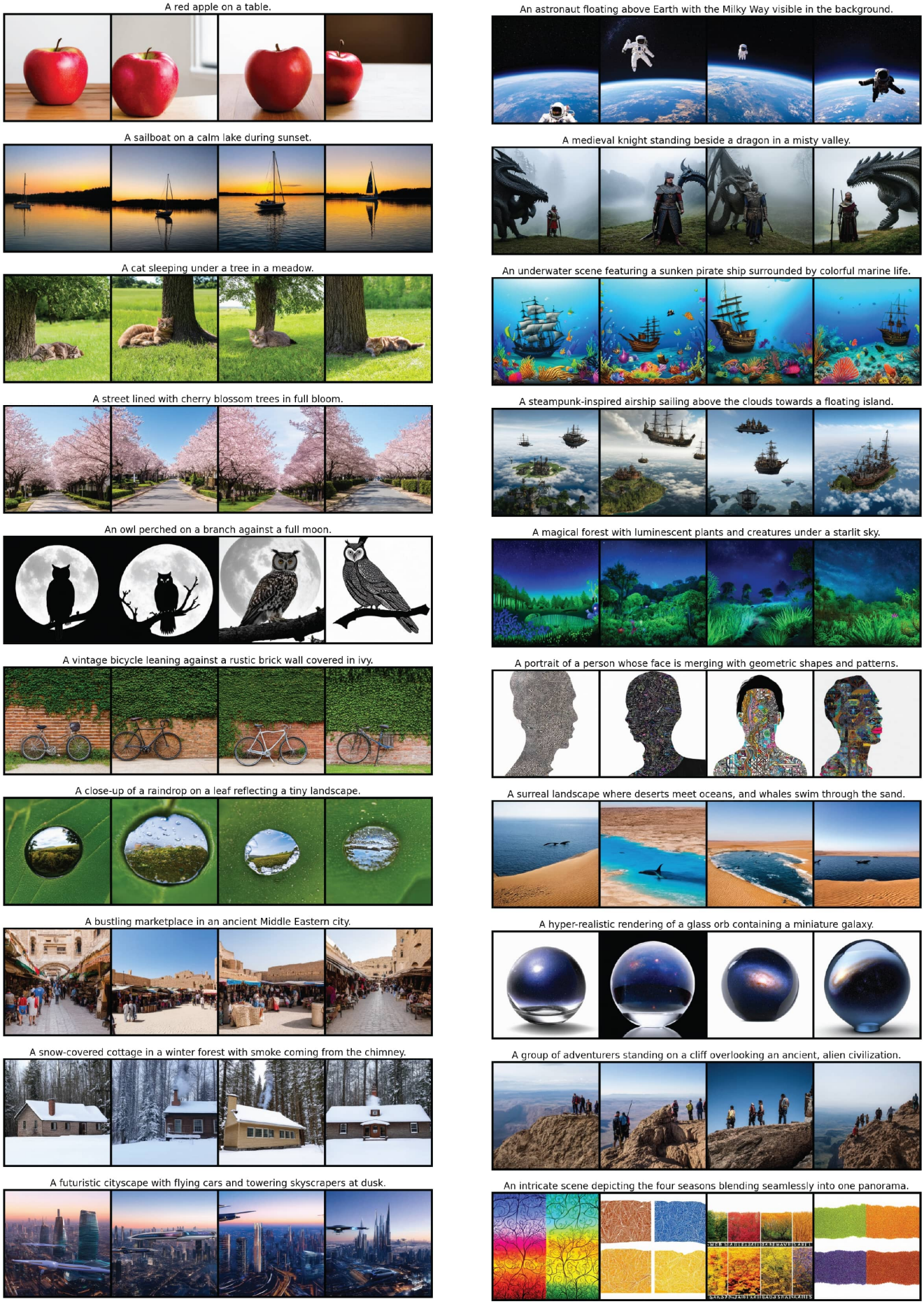


Figure 7. Additional text-to-image samples generated from the model.

References

[1] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pages 1692–1717. PMLR, 2023. 2, 4

[2] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 1

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1

[4] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 1

[5] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 1

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 1

[7] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention, 2024. 1

[8] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[9] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 1

[10] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 1